On the complexity of listing closed frequent subgraph patterns

Jan Ramon and Mostafa Haghir Chehreghani

Given a graph class $\mathcal{L}$, a matching operator $\leq$ (usually subgraph isomorphism), a multiset DB of elements of $\mathcal{L}$ (the database) and a treshold t, the problem $F(\mathcal{L}, \leq, DB, t)$ of frequent subgraph pattern mining is to list all elements P in $\mathcal{L}$ for which the frequency $freq(P, DB) = \#\{G \in DB \mid P \leq G\}$ is at least t. The problem parameter are the number of graphs in DB and the number of nodes in the largest graph in DB. The task of frequent subgraph pattern mining is a fundamental problem studied in the field of data mining. In previous work, Horvath and Ramon have studied the complexity of this listing problem. For the class of all trees, and for any monotone graph class for which the matching operator $\leq$ can be decided in polynomial time, the problem $F(\mathcal{L}, \leq, DB, t)$ can be solved with polynomial delay. On the other hand, in the general case, e.g. when $\mathcal{L}$ is the class of all graphs and $\leq$ is the subgraph isomorphism relation, $F(\mathcal{L}, \leq, DB, t)$ can not even be solved in output-polynomial time. Surprisingly, in some cases, such as for the class of graphs of bounded treewidth, the matching operator (subgraph isomorphism) is NP-hard while it is still possible to solve $F(\mathcal{L}, \leq, DB, t)$ in incremental polynomial time.

Because the number of frequent subgraph patterns may be huge in practice, several approaches have been considered to reduce the number of patterns without losing information. One way is to only remember closed patterns. Several notions of "closed pattern" have been considered. A subgraph pattern P is frequency-closed if each of its supergraphs have a strictly smaller frequency. A subgraph pattern P is embedding-closed if there does not exist a supergraph $C > P$ such that for each embedding $\varphi_P : P \rightarrow G$ of P into a graph $G \in DB$, this embedding can be extended into an embedding $\varphi_C \supset \varphi_P$ of C into G.

In this work, we study the complexity of mining frequent closed subgraph patterns, i.e. given a graph class $\mathcal{L}$, matching operator $\leq$, database DB, treshold t and notion of closed pattern cl, the problem $FC(\mathcal{L}, \leq, DB, t, cl)$ is to list all cl-closed patterns which have frequency at least t in DB. In particular, we show the following new results: i) if the matching operator $\leq$ is NP-hard, then $FC(\mathcal{L}, \leq, DB, t, cl)$ can not be solved in output-polynomial time. ii) if $\mathcal{L}$ is a monotone graph class and $\leq$ can be decided in polynomial time, then $FC(\mathcal{L}, \leq, DB, t, cl)$ can be solved with polynomial delay. iii) for trees, mining frequency-closed subtree patterns is not possible with polynomial delay (unless P=NP). We also provide upper bounds for the complexity of mining embedding-closed trees and graphs of bounded treewidth and degree.